Complex-Valued Restricted Boltzmann Machine for Speaker-Dependent Speech Parameterization from Complex Spectra

Complex-Valued Restricted Boltzmann Machine for Speaker-Dependent Speech Parameterization from Complex Spectra

Toru Nakashika, *Member, IEEE,*, Shinji Takaki, *Member, IEEE* and Junichi Yamagishi, *Senior Member, IEEE*

***Abstract*—This paper describes a novel energy-based probabilistic distribution that represents complex-valued data and explains how to apply it to direct feature extraction from complex-valued spectra. The proposed model, the complex-valued restricted Boltzmann machine (CRBM), is designed to deal with complex-valued visible units as an extension of the well-known restricted Boltzmann machine (RBM). Like the RBM, the CRBM learns the relationships between visible and hidden units without having connections between units in the same layer, which dramatically improves training efficiency by using Gibbs sampling or contrastive divergence (CD). Another important characteristic is that the CRBM also has connections between real and imaginary parts of each of the complex-valued visible units that help represent the data distribution in the complex domain. In speech signal processing, classification and generation features are often based on amplitude spectra (e.g., MFCC, cepstra, and mel-cepstra) even if they are calculated from complex spectra, and they ignore phase information. In contrast, the proposed feature extractor using the CRBM directly encodes the complex spectra (or another complex-valued representation of the complex spectra) into binary-valued latent features (hidden units). Since the visible-hidden connections are undirected, we can also recover (decode) the complex spectra from the latent features directly. Our speech representation experiments demonstrated that the CRBM outperformed other speech representation methods, such as methods using a conventional RBM, a mel-log spectrum approximate (MLSA) decoder, etc.**

***Index Terms*—restricted Boltzmann machine, deep learning, complex-valued representation, feature extraction, speech synthesis**

I. Introduction

DEEP LEARNING is one of the hottest recent topics in a wide range of research fields, such as artificial intelligence, machine learning, and signal processing that includes image classification, speech recognition, etc[1]. Many models have been proposed as deep learning tools; one of the most widely-used and famous models is a deep belief-net (DBN) [2] that stacks multiple restricted Boltzmann machines (RBMs) layer-by-layer. The RBM is a probabilistic model that consists of visible and hidden units and has often been used alone as a feature extractor, a generator, and as a classifier and pre-training scheme of deep neural networks. Many extensions of the RBM have been proposed for task specification [3], [4], [5], [6]. Although the RBM has been used in many tasks, the RBM traditionally identified visible units as either binary-valued or real-valued [2], [7], [8].

Representations based on the amplitude spectra of speech (such as MFCC, cepstra, and mel-cepstra) are traditionally used in speech signal processing as input features of speech recognition or output features of speech synthesis because the amplitude spectra are more effective and relevant to our auditory field for such tasks than phase spectra. Raw amplitude spectral representation can also be used [9], [10]. However, these features that include the amplitude spectra theoretically lack phase information, and single use of the amplitude-based features cannot completely recover the original complex spectra with reasonable computational resources easily, even when using the well-known Griffin-Lim algorithm [11]. Especially, the cepstral features are often inverted into a signal using typical vocoders, which assume that the spectrum has a minimum phase, even though a speech waveform is not always a minimum-phase signal. The well-known STRAIGHT [12] and the WORLD [13], which are speech transformation and representation methods in the spectral domain, are also based on the minimum-phase reconstruction. In [14], [15], [16], it is reported that the generated speech signals from direct waveform modification or synthesis are much more natural than those from methods that are based on phase reconstruction from amplitude spectra. Furthermore, there are many cases in other kinds of signal processing in which we have to deal with complex-valued actual data such as fMRI images, wireless signals, acoustic intensity, etc. Other machine learning models—that is, neural networks, Boltzmann machines, and non-negative matrix factorization (NMF) [17]—have their extensions proposed to represent complex-valued data [18], [19], [20]. Another example to model speech signals with both amplitude and phase is phase embedding features, such as complex cepstrum [21] and HMPD (harmonic model + phase distortion) [22].

In our previous work [23], we proposed an extended model of the RBM, namely the "complex-valued RBM (CRBM), " to represent complex-valued data in the RBM-based approach in particular. The CRBM includes three important characteristics. First, the CRBM has no connections across dimensions in the same layers but has connections between visible and hidden units like the RBMs. These restrictions make it exceedingly easy to estimate the parameters using Gibbs sampling or CD [2]. The "directional-unit Boltzmann machine (DUBM)" extension [19] has been also proposed to model complex-

T. Nakashika is with the Graduate School of Informatics and Engineering, the University of Electro-Communications, Tokyo, Japan e-mail: nakashika@uec.ac.jp

S. Takaki and J. Yamagishi are with National Institute of Informatics, Tokyo, Japan e-mail: takaki@nii.ac.jp, jyamagis@nii.ac.jp

Manuscript received April 19, 2005; revised December 27, 2012.

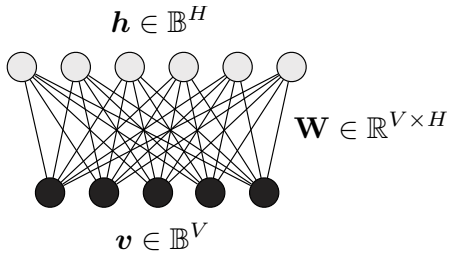
valued data. However, the DUBM has connections across dimensions, thus has difficulties in parameter estimation, unlike the CRBM. Second, unlike the conventional RBM, the CRBM restricts the connections between different visible units but still has connections between real and imaginary parts of each visible unit. Therefore, the CRBM represents the complex-valued data distribution more accurately than the RBMs, especially when there are correlations between the real and imaginary parts. Third, the CRBM represents the complex-valued visible units in a rectangular form that consists of real and imaginary components, while traditional representation methods of complex-valued data that include a DUBM are based on a polar form of phase and amplitude components. We can generate samples from the distribution straightforwardly in the CRBM. The conditional probability of visible units given hidden units forms a complex-normal distribution, which makes the real and imaginary components Gaussian-distributed. Therefore, the CRBM is not ideal for data that are not partially Gaussian-distributed. In [23], we showed that the CRBM sufficiently recovered the amplitude and phase components and the real and imaginary components in the speech representation experiments.

We also propose some improvements and learning techniques for the CRBM-based speech parameterization. First, we reduce the number of dimensions by feeding complex-valued visible features obtained by complex principal component analysis (CPCA) [24] into the CRBM instead of the raw complex spectra. Next, we employ the maximum likelihood parameter generation (MLPG) [25] to generate the trajectories of the CPCA features for better representation of speech sequences. Finally, we extend the Adam algorithm to deal with the complex-valued parameters (referred to as "complex Adam" or "CAdam"), which makes convergence of the model training faster than the steepest descent/ascent. In the experiments, we compare the performance of the improved CRBM method with other speech representation methods, such as the conventional RBM, the mel-log spectrum approximate (MLSA), etc.

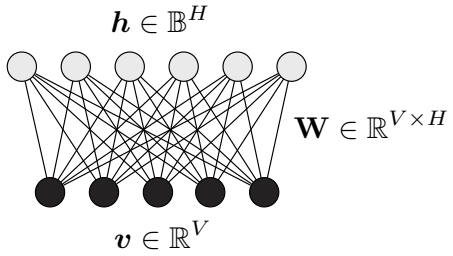
This paper is organized as follows. In Section 2, we briefly review the conventional RBM. In Section 3, we present our proposed model, CRBM. In Section 4, we present improvement methods for the CRBM using the CPCA. In Section 5, we propose a complex-valued sequence generation method based on MLPG. In Section 6, we show our experimental results. In Section 7, we conclude our findings.

II. Preliminary

A restricted Boltzmann machine (RBM) is one of the most widely used energy-based models and is convenient for representing latent features that cannot be observed but surely exist in the background. The Bernoulli-Bernoulli RBM (BB-RBM) was originally introduced by Freund *et. al* [7]. It defines the distribution of binary-valued visible variables $\boldsymbol{v} \in \mathbb{B}^V$ and binary-valued hidden (latent) variables $\boldsymbol{h} \in \mathbb{R}^H$ with their undirected real-valued connection weights $\mathbf{W} \in \mathbb{R}^{V \times H}$, as shown in Fig. 1 (a) where V and H are the numbers of dimensions in their respective visible and hidden units and



(a) Bernoulli-Bernoulli RBM



(b) Gaussian-Bernoulli RBM

Fig. 1: Graphical representation of conventional RBMs.

$\mathbb{B} \triangleq \{0, 1\}$ indicates the binary set. The RBM was later extended to deal with real-valued data known as a "Gaussian-Bernoulli RBM (GB-RBM)" [8], as shown in Fig. 1 (b). However, it has been reported that there were some difficulties with the original GB-RBM because of the unstable training of the parameters. Later, Cho *et al.* [26] proposed an improved learning method for a GB-RBM to overcome the difficulties. In the remainder of this paper, we refer to this improved GB-RBM as just an RBM unless otherwise stated. In the modeling using an RBM, the joint probability $p(\boldsymbol{v}, \boldsymbol{h})$ of real-valued visible units $\boldsymbol{v}$ and binary-valued hidden units $\boldsymbol{h}$ is defined as follows:

$$p(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \frac{1}{U(\boldsymbol{\theta})} e^{-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})} \tag{1}$$

$$E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta}) = \frac{1}{2} \boldsymbol{v}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{v} - \boldsymbol{b}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{v} - \boldsymbol{c}^\top \boldsymbol{h} - \boldsymbol{v}^\top \boldsymbol{\Sigma}^{-1} \mathbf{W} \boldsymbol{h} \tag{2}$$

$$U(\boldsymbol{\theta}) = \int \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v}, \boldsymbol{h}; \boldsymbol{\theta})} d\boldsymbol{v} \tag{3}$$

where $\boldsymbol{\theta} = \{\boldsymbol{b}, \boldsymbol{c}, \mathbf{W}, \boldsymbol{\sigma}\}$ indicates a set of parameters that contains bias parameters of the visible units $\boldsymbol{b} \in \mathbb{R}^V$, bias parameters of the hidden units $\boldsymbol{c} \in \mathbb{R}^H$, the connection weight parameters between visible-hidden units $\mathbf{W} \in \mathbb{R}^{V \times H}$, and the standard deviation parameters associated with the dimension independent Gaussian visible units $\boldsymbol{\sigma} \in \mathbb{R}^V$ that define $\boldsymbol{\Sigma} \triangleq \mathrm{diag}(\boldsymbol{\sigma}^2)$ (the function $\mathrm{diag}(\cdot)$ returns a diagonal matrix whose diagonal vector is the argument, and $\cdot^2$ indicates the element-wise square operation). The parameters $\boldsymbol{\theta}$ are often estimated using the maximum likelihood (ML) and the gradient descent/ascent given the training set $D \ni \boldsymbol{v}$. The partial gradients of the parameters to the expected log

likelihood:

$$L(\boldsymbol{\theta}) = \mathbb{E}_D[\log p(\boldsymbol{v};\boldsymbol{\theta})] = \mathbb{E}_D[\log \sum_{\boldsymbol{h}} p(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta})] \quad (4)$$

can be calculated as:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \langle -\frac{\partial E}{\partial \boldsymbol{\theta}} \rangle_{\mathrm{data}} - \langle -\frac{\partial E}{\partial \boldsymbol{\theta}} \rangle_{\mathrm{model}}, \quad (5)$$

where $\langle\cdot\rangle_{\mathrm{data}}$ and $\langle\cdot\rangle_{\mathrm{model}}$ indicate expectations of the training data and the inner model, respectively. Although exact calculation of the inner model has an order of 2^{V+H}, the expectation value can be approximated using the Gibbs sampling, or more efficiently, the contrastive divergence (CD) [2]. From the definition of RBM in Eqs. (1), (2), and (3), the conditional probabilities $p(\boldsymbol{v}|\boldsymbol{h})$ and $p(\boldsymbol{h}|\boldsymbol{v})$ form quite simple distributions as:

$$p(\boldsymbol{v}|\boldsymbol{h}) = \mathcal{N}(\boldsymbol{v};\boldsymbol{b}+\mathbf{W}\boldsymbol{h},\boldsymbol{\Sigma}) \quad (6)$$

$$p(\boldsymbol{h}|\boldsymbol{v}) = \mathcal{B}(\boldsymbol{h};\boldsymbol{f}(\boldsymbol{c}+\mathbf{W}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{v})) \quad (7)$$

where $\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\mathcal{B}(\cdot;\boldsymbol{\pi})$, and $\boldsymbol{f}(\cdot)$ indicate the multivariate Gaussian distribution with the mean $\boldsymbol{\mu}$ and the variance matrix $\boldsymbol{\Sigma}$, the multi-dimensional Bernoulli distribution with the success probabilities $\boldsymbol{\pi}$, and an element-wise sigmoid function, respectively. As Eqs. (6) and (7) indicates, we can easily compute the iteration of drawing samples $\boldsymbol{h}$ given $\boldsymbol{v}$, and $\boldsymbol{v}$ given $\boldsymbol{h}$, which is used in Gibbs sampling or CD. The same is true for BB-RBM. In the case of BB-RBM, the conditional probabilities $p(\boldsymbol{v}|\boldsymbol{h})$ and $p(\boldsymbol{h}|\boldsymbol{v})$ turn into the following:

$$p(\boldsymbol{v}|\boldsymbol{h}) = \mathcal{B}(\boldsymbol{v};\boldsymbol{f}(\boldsymbol{b}+\mathbf{W}\boldsymbol{h}) \quad (8)$$

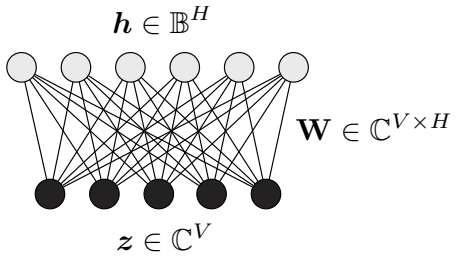
$$p(\boldsymbol{h}|\boldsymbol{v}) = \mathcal{B}(\boldsymbol{h};\boldsymbol{f}(\boldsymbol{c}+\mathbf{W}^\top\boldsymbol{v})) \quad (9)$$

under the energy function:

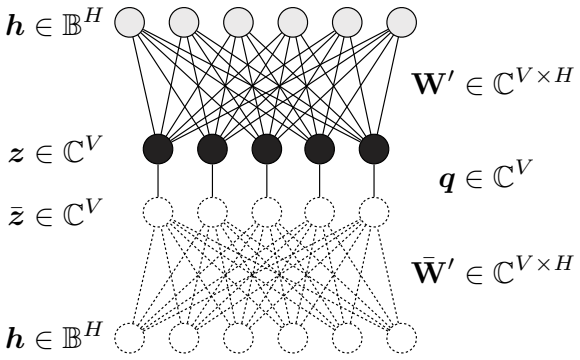
$$E(\boldsymbol{v},\boldsymbol{h};\boldsymbol{\theta}) = -\boldsymbol{b}^\top\boldsymbol{v} - \boldsymbol{c}^\top\boldsymbol{h} - \boldsymbol{v}^\top\mathbf{W}\boldsymbol{h}. \quad (10)$$

III. Complex-valued RBM

Conventional RBMs assume that data is either binary-valued or real-valued. Therefore, complex-valued data should not be fed into conventional RBMs directly because the conditional probability of visible units specifies binary- or real-valued variables, as Eqs. (6) and (8) indicate. In other words, the conditional probability of visible units should specify complex-valued variables in order to feed complex-valued variables into the model. In our approach, we define an extension of the RBM that feeds complex-valued data and forms the conditional probability of visible units as complex normal distribution [27]. A real-valued cost function (the likelihood) is still used in parameter estimation for an extended RBM—namely a complex-valued RBM (CRBM)—because the probability distribution is real-valued. Like conventional RBMs, the CRBM consists of two layers: complex-valued visible units $\boldsymbol{z}$ and binary-valued hidden units $\boldsymbol{h}$ with undirected connection weights $\mathbf{W}$, as shown in Fig. 2 (a). Furthermore, in the CRBM, we give a "restriction" where there are no connections between visible units or hidden units, which enables easy estimation of parameters just as an RBM does. However, we allow the model to have connections between the real and imaginary parts in order to capture the relationships between the real and imaginary parts of each complex-valued visible unit.



(a) Complex-valued RBM (CRBM)



(b) Another representation of CRBM

Fig. 2: Graphical representation of a complex-valued RBM.

A. Definition

Based on the above discussion, we formulate the CRBM as the joint probability of complex-valued visible units $\boldsymbol{z} \in \mathbb{C}^V$ and binary-valued hidden units $\boldsymbol{h} \in \mathbb{B}^H$ with an energy function that defines the relations between $\boldsymbol{z}$ and $\boldsymbol{h}$, where V and H are the numbers of visible and hidden units. This is given by:

$$p(\boldsymbol{z},\boldsymbol{h};\boldsymbol{\theta}) = \frac{1}{U(\boldsymbol{\theta})} e^{-E(\boldsymbol{z},\boldsymbol{h};\boldsymbol{\theta})} \quad (11)$$

$$\begin{aligned} E(\boldsymbol{z},\boldsymbol{h};\boldsymbol{\theta}) =& \frac{1}{2}\begin{bmatrix}\boldsymbol{z}\\ \bar{\boldsymbol{z}}\end{bmatrix}^H \boldsymbol{\Phi}^{-1}\begin{bmatrix}\boldsymbol{z}\\ \bar{\boldsymbol{z}}\end{bmatrix} \\ & - \begin{bmatrix}\boldsymbol{b}\\ \bar{\boldsymbol{b}}\end{bmatrix}^H \boldsymbol{\Phi}^{-1}\begin{bmatrix}\boldsymbol{z}\\ \bar{\boldsymbol{z}}\end{bmatrix} - 2\boldsymbol{c}^\top\boldsymbol{h} \\ & - \begin{bmatrix}\boldsymbol{z}\\ \bar{\boldsymbol{z}}\end{bmatrix}^H \boldsymbol{\Phi}^{-1}\begin{bmatrix}\mathbf{W}\\ \bar{\mathbf{W}}\end{bmatrix}\boldsymbol{h} \end{aligned} \quad (12)$$

$$U(\boldsymbol{\theta}) = \int \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{z},\boldsymbol{h};\boldsymbol{\theta})} d\boldsymbol{z}, \quad (13)$$

where $\bar{\cdot}$ denotes complex-conjugate and $\cdot^H$ denotes Hermitian-transpose. $\boldsymbol{b} \in \mathbb{C}^V$, $\boldsymbol{c} \in \mathbb{R}^H$, and $\mathbf{W} \in \mathbb{C}^{V\times H}$ are bias parameters of the visible units and the hidden units, and the *biased* connection weights between visible and hidden units, respectively. The second and third terms of the energy function E in Eq. (12) indicate the potential of individual visible and hidden units, and the fourth term indicates the pairwise potential (intensity of relationships) between visible and hidden units. The first term is put to make the visible units

complex-Gaussian-distributed.

In order to make the restrictions that the model has no connections across dimensions of visible units but has connections between real and imaginary parts of each dimension, the extended covariance matrix $\mathbf{\Phi}$ consists of a covariance matrix $\mathbf{\Gamma}$ and a pseudo-covariance matrix $\mathbf{C}$—both of which are diagonal matrices—as

$$\mathbf{\Phi} \triangleq \begin{bmatrix} \mathbf{\Gamma} & \mathbf{C} \\ \mathbf{C}^H & \mathbf{\Gamma}^H \end{bmatrix} \tag{14}$$

and

$$\begin{aligned} \mathbf{\Gamma} &\triangleq \mathrm{diag}(\boldsymbol{\gamma}), \quad \boldsymbol{\gamma} \in \mathbb{R}+^V \\ \mathbf{C} &\triangleq \mathrm{diag}(\boldsymbol{\delta}), \quad \boldsymbol{\delta} \in \mathbb{C}^V \end{aligned} \tag{15}$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are variance and pseudo-variance parameters of the complex-valued visible units, respectively. To summarize, the set of parameters of the CRBM is $\boldsymbol{\theta} = \{\boldsymbol{b}, \boldsymbol{c}, \mathbf{W}, \boldsymbol{\gamma}, \boldsymbol{\delta}\}$.

Introducing auxiliary precision vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ defined as

$$\boldsymbol{p} \triangleq \frac{\boldsymbol{\gamma}}{\boldsymbol{\gamma}^2 - |\boldsymbol{\delta}|^2} \in \mathbb{R}^V \tag{16}$$

$$\boldsymbol{q} \triangleq -\frac{\boldsymbol{\delta}}{\boldsymbol{\gamma}^2 - |\boldsymbol{\delta}|^2} \in \mathbb{C}^V \tag{17}$$

where the fraction bar denotes element-wise division, we can rewrite the energy function in Eq. (12) as follows:

$$\begin{aligned} &E(\boldsymbol{z}, \boldsymbol{h}; \boldsymbol{\theta}) = \\ &\boldsymbol{z}^H \mathrm{diag}(\boldsymbol{p}) \boldsymbol{z} + \Re(\boldsymbol{z}^H \mathrm{diag}(\boldsymbol{q}) \bar{\boldsymbol{z}}) - 2\Re(\boldsymbol{z}^H \mathrm{diag}(\boldsymbol{p}) \boldsymbol{b}) \\ &- 2\Re(\boldsymbol{z}^H \mathrm{diag}(\boldsymbol{q}) \bar{\boldsymbol{b}}) - 2\boldsymbol{c}^\top \boldsymbol{h} - 2\Re(\boldsymbol{z}^H \mathrm{diag}(\boldsymbol{p}) \mathbf{W}) \boldsymbol{h} \\ &- 2\Re(\boldsymbol{z}^H \mathrm{diag}(\boldsymbol{q}) \bar{\mathbf{W}}) \boldsymbol{h}, \end{aligned} \tag{18}$$

which confirms that 1) the above energy function E and the probability distribution are real-valued and that 2) there are connections between the complex-valued visible units and their conjugates for each dimension but there are no connections between different dimensions.

Furthermore, when we use unbiased parameters:

$$\boldsymbol{b}' \triangleq \mathrm{diag}(\boldsymbol{p}) \boldsymbol{b} + \mathrm{diag}(\boldsymbol{q}) \bar{\boldsymbol{b}} \tag{19}$$

$$\mathbf{W}' \triangleq \mathrm{diag}(\boldsymbol{p}) \mathbf{W} + \mathrm{diag}(\boldsymbol{q}) \bar{\mathbf{W}}, \tag{20}$$

the energy function E becomes

$$\begin{aligned} &E(\boldsymbol{z}, \boldsymbol{h}; \boldsymbol{\theta}) = \\ &\frac{1}{2} \boldsymbol{z}^H \mathrm{diag}(\boldsymbol{p}) \boldsymbol{z} + \frac{1}{2} \bar{\boldsymbol{z}}^H \mathrm{diag}(\boldsymbol{p}) \bar{\boldsymbol{z}} + \boldsymbol{z}^H \mathrm{diag}(\boldsymbol{q}) \bar{\boldsymbol{z}} \\ &+ \bar{\boldsymbol{z}}^H \mathrm{diag}(\bar{\boldsymbol{q}}) \boldsymbol{z} - \boldsymbol{z}^H \boldsymbol{b}' - \bar{\boldsymbol{z}}^H \bar{\boldsymbol{b}}' - 2\boldsymbol{c}^\top \boldsymbol{h} \\ &- \boldsymbol{z}^H \mathbf{W}' \boldsymbol{h} - \bar{\boldsymbol{z}}^H \bar{\mathbf{W}}' \boldsymbol{h}, \end{aligned} \tag{21}$$

which indicates that $\boldsymbol{z}$ and $\bar{\boldsymbol{z}}$ are symmetric to each other, as shown in Figure 2 (b).

From the above definition, the conditional probabilities $p(\boldsymbol{z}|\boldsymbol{h})$ and $p(\boldsymbol{h}|\boldsymbol{z})$ can be derived as follows:

$$p(\boldsymbol{z}|\boldsymbol{h}) = \mathcal{CN}(\boldsymbol{z}; \boldsymbol{b} + \mathbf{W}\boldsymbol{h}, \mathbf{\Gamma}, \mathbf{C}) \tag{22}$$

$$p(\boldsymbol{h}|\boldsymbol{z}) = \mathcal{B}(\boldsymbol{h}; \boldsymbol{f}(2\boldsymbol{c} + 2\Re(\mathbf{W}'^H \boldsymbol{z}))) \tag{23}$$

where $\mathcal{CN}(\cdot; \boldsymbol{\mu}, \mathbf{\Gamma}, \mathbf{C})$ is a multivariate complex normal distribution [27], a mean vector $\boldsymbol{\mu}$, a covariance matrix $\mathbf{\Gamma}$, and a pseudo-covariance matrix $\mathbf{C}$:

$$\begin{aligned} p(\boldsymbol{z}) &= \frac{1}{\pi^D \sqrt{\det(\mathbf{\Gamma}) \det(\mathbf{Q})}} \\ &\cdot \exp\left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{z} - \boldsymbol{\mu} \\ \bar{\boldsymbol{z}} - \bar{\boldsymbol{\mu}} \end{bmatrix}^H \begin{bmatrix} \mathbf{\Gamma} & \mathbf{C} \\ \mathbf{C}^H & \mathbf{\Gamma}^H \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{z} - \boldsymbol{\mu} \\ \bar{\boldsymbol{z}} - \bar{\boldsymbol{\mu}} \end{bmatrix} \right\} \end{aligned} \tag{24}$$

$$\mathbf{Q} = \bar{\mathbf{\Gamma}} - \mathbf{C}^H \mathbf{\Gamma}^{-1} \mathbf{C}. \tag{25}$$

B. Parameter estimation

To estimate the model parameters $\boldsymbol{\theta}$ of the CRBM, we employed the complex-valued gradient method. In this approach, the parameters $\boldsymbol{\theta}$ are estimated so as to maximize the expected log-likelihood L of the complex-valued training data set $D \ni \boldsymbol{z}$:

$$L(\boldsymbol{\theta}) = \mathbb{E}_D[\log p(\boldsymbol{z}; \boldsymbol{\theta})] \tag{26}$$

$$= \mathbb{E}_D[\log \sum_{\boldsymbol{h}} p(\boldsymbol{z}, \boldsymbol{h}; \boldsymbol{\theta})] \tag{27}$$

$$= \mathbb{E}_D[\log \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{z}, \boldsymbol{h}; \boldsymbol{\theta})}] - \log \int \sum_{\tilde{\boldsymbol{h}}} e^{-E(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{h}}; \boldsymbol{\theta})} d\tilde{\boldsymbol{z}}. \tag{28}$$

The complex-valued gradient ascend iteratively updates each parameter as:

$$\boldsymbol{\theta}^{(l+1)} \leftarrow \boldsymbol{\theta}^{(l)} + \boldsymbol{g}^{(l)}(\frac{\partial L}{\partial \boldsymbol{\theta}}) \tag{29}$$

where $\boldsymbol{\theta}^{(l)}$ indicates the parameters and $\boldsymbol{g}^{(l)}$ indicates the complex-valued gradient at the l-th iteration. One of the simplest gradient functions is the complex-valued steepest ascent (CSA) [28], [29], which is:

$$\boldsymbol{g}^{(l)}(\frac{\partial L}{\partial \boldsymbol{\theta}}) = 2\alpha \frac{\partial L}{\partial \bar{\boldsymbol{\theta}}} \tag{30}$$

where $\alpha \in \mathbb{C}, \Re(\alpha) > 0$ is a complex-valued learning rate. A simple CSA is not be suitable for a large amount of speech training data due to the slow convergence speed. Therefore, we propose another, more efficient learning method, the complex-valued adaptive momentum (CAdam), which is motivated by the real-valued Adam algorithm [30]. In the CAdam, we introduce auxiliary parameters $\boldsymbol{m}^{(n)}$ and $\boldsymbol{v}^{(n)}$ and update the parameters as:

$$\boldsymbol{m}^{(l)} = \beta_1 \boldsymbol{m}^{(l-1)} + (1 - \beta_1) \nabla_{\bar{\boldsymbol{\theta}}} L \tag{31}$$

$$\boldsymbol{v}^{(l)} = \beta_2 \boldsymbol{v}^{(l-1)} + (1 - \beta_2) |\nabla_{\bar{\boldsymbol{\theta}}} L|^2 \tag{32}$$

$$\Delta \boldsymbol{\theta}^{(l)} = 2\alpha \frac{1 - \beta_2^l}{1 - \beta_1^l} \frac{\boldsymbol{m}^{(l)}}{\boldsymbol{v}^{(l)}}, \tag{33}$$

where $\beta_1, \beta_2 \in \mathbb{R}, 0 < \beta_1, \beta_2 < 1$, and $\alpha \in \mathbb{C}, \Re(\alpha) > 0$.

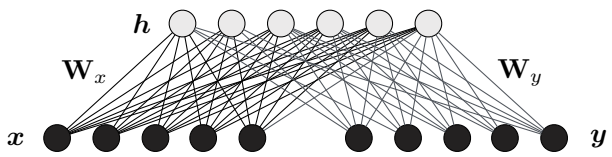
Calculating partial gradients to the parameters $\boldsymbol{\theta}$, we obtain:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \langle -\frac{\partial E}{\partial \boldsymbol{\theta}} \rangle_{\mathrm{data}} - \langle -\frac{\partial E}{\partial \boldsymbol{\theta}} \rangle_{\mathrm{model}}, \tag{34}$$

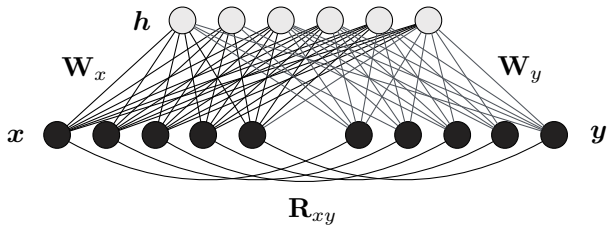
where the complex-valued partial gradients here indicate the Wirtinger derivatives:

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{1}{2} \left(\frac{\partial L}{\partial \Re(\boldsymbol{\theta})} - i \frac{\partial L}{\partial \Im(\boldsymbol{\theta})} \right) \tag{35}$$

$$\frac{\partial L}{\partial \bar{\boldsymbol{\theta}}} = \frac{1}{2} \left(\frac{\partial L}{\partial \Re(\boldsymbol{\theta})} + i \frac{\partial L}{\partial \Im(\boldsymbol{\theta})} \right). \tag{36}$$



(a) Concatenating GB-RBM without dimension-tied restriction



(b) Concatenating GB-RBM with dimension-tied restriction (with the same restriction as CRBM)

Fig. 3: Representing complex-valued data using a concatenating GB-RBM (a) without and (b) with dimension-tied restriction.

The negative partial gradients of the energy function with respect to each parameter $-\frac{\partial E}{\partial \boldsymbol{\theta}}$ can be further derived as:

$$-\frac{\partial E}{\partial \boldsymbol{b}} = \mathrm{diag}(\boldsymbol{p})\bar{\boldsymbol{z}} + \mathrm{diag}(\bar{\boldsymbol{q}})\boldsymbol{z} \tag{37}$$

$$-\frac{\partial E}{\partial \boldsymbol{c}} = \boldsymbol{h} \tag{38}$$

$$-\frac{\partial E}{\partial \mathbf{W}} = (\mathrm{diag}(\boldsymbol{p})\bar{\boldsymbol{z}} + \mathrm{diag}(\bar{\boldsymbol{q}})\boldsymbol{z})\boldsymbol{h}^\top \tag{39}$$

$$-\frac{\partial E}{\partial \boldsymbol{\gamma}} = (\boldsymbol{p}^2 + |\boldsymbol{q}|^2) \circ \frac{\partial E}{\partial \boldsymbol{p}} + 2\Re(\boldsymbol{p} \circ \boldsymbol{q} \circ \frac{\partial E}{\partial \boldsymbol{q}}) \tag{40}$$

$$-\frac{\partial E}{\partial \boldsymbol{\delta}} = \boldsymbol{p}^2 \circ \frac{\partial E}{\partial \boldsymbol{q}} + \bar{\boldsymbol{q}}^2 \circ \frac{\partial E}{\partial \bar{\boldsymbol{q}}} + 2\boldsymbol{p} \circ \bar{\boldsymbol{q}} \circ \frac{\partial E}{\partial \boldsymbol{p}}, \tag{41}$$

where $\circ$ denotes an element-wise product and $|\cdot|$ denotes the absolute, and

$$\frac{\partial E}{\partial \boldsymbol{p}} = \frac{1}{2}|\boldsymbol{z}|^2 - \Re(\boldsymbol{z} \circ (\bar{\boldsymbol{b}} + \bar{\mathbf{W}}\boldsymbol{h})) \tag{42}$$

$$\frac{\partial E}{\partial \boldsymbol{q}} = \frac{1}{2}\bar{\boldsymbol{z}}^2 - \bar{\boldsymbol{z}} \circ (\bar{\boldsymbol{b}} + \bar{\mathbf{W}}\boldsymbol{h}). \tag{43}$$

The gradients of variance and pseudo variance tend to be larger than those of the other parameters. For stable training, we replace the parameters as $\boldsymbol{\gamma} \triangleq e^{\boldsymbol{r}}$ and $\boldsymbol{\delta} \triangleq e^{\boldsymbol{s}}$ and update using the gradients of $\boldsymbol{r}$ and $\boldsymbol{s}$ in a manner similar to the improved GB-RBM [26].

The second term on the right-hand side in Eq. (34) usually requires a high computational cost. However, because of the restrictions of the CRBM, the second term can be efficiently approximated using Gibbs sampling or CD [2] in a way similar to conventional RBMs.

C. Relationships with complex representation using GB-RBM

We can also represent a complex-valued vector $\boldsymbol{z} = \boldsymbol{x} + i\boldsymbol{y}$ (where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^V$) in the real space using a conventional GB-RBM that feeds the double-sized concatenated vector $\boldsymbol{z}' = [\boldsymbol{x}^\top \ \boldsymbol{y}^\top]^\top \in \mathbb{R}^{2V}$ as:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{h}} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h}) \tag{44}$$

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h}) = \frac{1}{U} e^{-E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h})} \tag{45}$$

$$\begin{aligned} E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h}) = &\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}_x^{-1}\boldsymbol{x} + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{\Sigma}_y^{-1}\boldsymbol{y} \\ &- \boldsymbol{b}_x^\top \boldsymbol{\Sigma}_x^{-1}\boldsymbol{x} - \boldsymbol{b}_y^\top \boldsymbol{\Sigma}_y^{-1}\boldsymbol{y} - \boldsymbol{c}^\top \boldsymbol{h} \\ &- \boldsymbol{x}^\top \boldsymbol{\Sigma}_x^{-1}\mathbf{W}_x\boldsymbol{h} - \boldsymbol{y}^\top \boldsymbol{\Sigma}_y^{-1}\mathbf{W}_y\boldsymbol{h} \end{aligned} \tag{46}$$

$$U = \int\int \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h})} d\boldsymbol{x} d\boldsymbol{y}, \tag{47}$$

where $\boldsymbol{\Sigma}_x = \mathrm{diag}(\boldsymbol{\sigma}_x^2), \boldsymbol{\Sigma}_y = \mathrm{diag}(\boldsymbol{\sigma}_y^2)$, and we decompose the GB-RBM parameters as $\boldsymbol{b} = [\boldsymbol{b}_x^\top \ \boldsymbol{b}_y^\top]^\top$, $\boldsymbol{\sigma} = [\boldsymbol{\sigma}_x^\top \ \boldsymbol{\sigma}_y^\top]^\top$, $\mathbf{W} = [\mathbf{W}_x^\top \ \mathbf{W}_y^\top]^\top$ in Eqs. (1), (2), and (3). Fig. 3 (a) depicts this concatenating representation of the complex-valued data using the GB-RBM. For example, the negative partial differentials of the real and imaginary parts of the bias parameters to the energy function in this representation can be derived as:

$$-\frac{\partial E}{\partial \boldsymbol{b}_x} = \frac{\boldsymbol{x}}{\boldsymbol{\sigma}_x^2} \tag{48}$$

$$-\frac{\partial E}{\partial \boldsymbol{b}_y} = \frac{\boldsymbol{y}}{\boldsymbol{\sigma}_y^2}. \tag{49}$$

On the other hand, when we put $\boldsymbol{z} = \boldsymbol{x} + i\boldsymbol{y}$, $\boldsymbol{b} = \boldsymbol{b}^R + i\boldsymbol{b}^I$, $\mathbf{W} = \mathbf{W}^R + i\mathbf{W}^I$, $\boldsymbol{q} = \boldsymbol{q}^R + i\boldsymbol{q}^I$ where $\boldsymbol{b}^R, \boldsymbol{b}^I \in \mathbb{R}^V$, $\mathbf{W}^R, \mathbf{W}^I \in \mathbb{R}^{V \times H}$, $\boldsymbol{q}^R, \boldsymbol{q}^I \in \mathbb{R}^V$, we can rewrite the energy function of CRBM in Eq. (12) as:

$$\begin{aligned} E = &\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}_x^{-1}\boldsymbol{x} + \boldsymbol{x}^\top \mathbf{R}_{xy}\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{\Sigma}_y^{-1}\boldsymbol{y} \\ &- \boldsymbol{b}_x^\top \boldsymbol{\Sigma}_x^{-1}\boldsymbol{x} - \boldsymbol{b}_y^\top \boldsymbol{\Sigma}_y^{-1}\boldsymbol{y} - \boldsymbol{c}^\top \boldsymbol{h} \\ &- \boldsymbol{x}^\top \boldsymbol{\Sigma}_x^{-1}\mathbf{W}_x\boldsymbol{h} - \boldsymbol{y}^\top \boldsymbol{\Sigma}_y^{-1}\mathbf{W}_y\boldsymbol{h} \end{aligned}, \tag{50}$$

where we introduce

$$\boldsymbol{\Sigma}_x = \mathrm{diag}(\frac{1}{2(\boldsymbol{p} + \boldsymbol{q}^R)}) \tag{51}$$

$$\boldsymbol{\Sigma}_y = \mathrm{diag}(\frac{1}{2(\boldsymbol{p} - \boldsymbol{q}^R)}) \tag{52}$$

$$\mathbf{R}_{xy} = \mathrm{diag}(2\boldsymbol{q}^I) \tag{53}$$

$$\boldsymbol{b}_x = \boldsymbol{b}^R + \frac{\boldsymbol{q}^I}{\boldsymbol{p} + \boldsymbol{q}^R} \circ \boldsymbol{b}^I \tag{54}$$

$$\boldsymbol{b}_y = \boldsymbol{b}^I + \frac{\boldsymbol{q}^I}{\boldsymbol{p} - \boldsymbol{q}^R} \circ \boldsymbol{b}^R \tag{55}$$

$$\mathbf{W}_x = \mathbf{W}^R + \mathrm{diag}(\frac{\boldsymbol{q}^I}{\boldsymbol{p} + \boldsymbol{q}^R})\mathbf{W}^I \tag{56}$$

$$\mathbf{W}_y = \mathbf{W}^I + \mathrm{diag}(\frac{\boldsymbol{q}^I}{\boldsymbol{p} - \boldsymbol{q}^R})\mathbf{W}^R. \tag{57}$$

Comparing the energy functions in Eqs. (46) and (50), the latter energy function includes a cross term ($\boldsymbol{x}^\top \mathbf{R}_{xy}\boldsymbol{y}$) between $\boldsymbol{x}$ and $\boldsymbol{y}$ while the former energy function does not. In the case of modeling complex values, the main difference between the CRBM and the conventional RBM is the connections between the real and imaginary parts for each dimension with the weights $\boldsymbol{r}_{xy} = 2\boldsymbol{q}^I$, as in Fig. 3 (b), where $\boldsymbol{r}_{xy}$ is the diagonal vector of $\mathbf{R}_{xy}$.

Another difference is the gradients. In the GB-RBM representation, the gradients regarding the real and imaginary parts of the bias of visible units, for example, are calculated independently of each other as Eqs. (48) and (49) indicate, while the gradients of the bias of visible units in the CRBM representation are calculated using both the real and imaginary terms as Eq. (37) indicates. This will make the model convergence better than the GB-RBM.

IV. Complex spectra compression using CPCA

The aim of this paper is to represent trajectories of complex-valued speech spectra. In general, the number of dimensions of the raw complex spectra tends to be large (e.g., when analyzing speech with the window length of $1,024$, the complex spectra has the dimensions of 513), which makes it difficult to use dynamic features or segment features as input for the model due to the sizable number of parameters. Therefore, we reduced the dimensions using complex principal component analysis (CPCA) [24] in this paper.

Letting $\boldsymbol{o}_t$ be the complex spectra at the frame t, the complex-valued features $\boldsymbol{z}_t$ whose dimensions are reduced to P using CPCA calculated as:

$$\boldsymbol{z}_t = \boldsymbol{\Lambda}_{1:P}^{-\frac{1}{2}} \mathbf{U}_{:,1:P}^H \boldsymbol{o}_t, \tag{58}$$

where $\boldsymbol{\Lambda}_{1:P}^{-\frac{1}{2}}$ and $\mathbf{U}_{:,1:P}$ indicate a diagonal matrix where the diagonal elements are the inverse of the top P eigenvalues of the empirical covariance matrix and a complex matrix whose columns are the complex eigenvectors corresponding to the eigenvalues. Conversely, when we recover the complex spectra $\boldsymbol{o}_t$ from $\boldsymbol{z}_t$, we just calculate the inversion as:

$$\boldsymbol{o}_t = \mathbf{U}_{:,1:P} \boldsymbol{\Lambda}_{1:P}^{\frac{1}{2}} \boldsymbol{z}_t. \tag{59}$$

In our speech modeling experiments that will be discussed later, we used the concatenated features $\boldsymbol{Z}_t \triangleq [\boldsymbol{z}_t^H \ \Delta \boldsymbol{z}_t^H]^H$ as visible units in CRBM, where $\boldsymbol{z}_t^H$ was calculated using the CPCA with the degree of $P = 40$ from the complex spectra analyzed with the window length of 256, and their dynamics $\Delta \boldsymbol{z}_t^H$ were calculated as $0.5\boldsymbol{z}_{t+1} - 0.5\boldsymbol{z}_{t-1}$. The total dimensions of visible units in CRBM were $I = 80$. In these experiments, the CRBM was trained so as to maximize the expected likelihood $L(\boldsymbol{\theta}) = \mathbb{E}[\log p(\boldsymbol{Z}; \boldsymbol{\theta})]$ of the concatenated feature set.

V. Complex-valued sequence generation based on MLPG

When we apply the CRBM to represent speech spectra, we need further improvements to compare it to other speech feature extraction methods. In this section, we present improved methods of dealing with trajectory modeling.

In our first work on the CRBM [23], we probabilistically encoded complex-valued visible units $\boldsymbol{z}_t$ into binary values by calculating the expectations of hidden units as $\hat{\boldsymbol{h}}_t \triangleq \mathbb{E}[p(\boldsymbol{h}_t|\boldsymbol{z}_t)]$ and inversely decoded (recovered) them from $\hat{\boldsymbol{h}}_t$ by calculating the expectations of visible units as $\hat{\boldsymbol{z}}_t \triangleq \mathbb{E}[p(\boldsymbol{z}_t|\hat{\boldsymbol{h}}_t)] = \boldsymbol{b} + \mathbf{W}\hat{\boldsymbol{h}}_t$ frame-by-frame. However, speech signals are sequences; there are correlations between adjacent frames of speech. In this paper, we employ trajectory modeling and sequence generation instead of frame-wise modeling. Our method efficiently recovers complex-valued visible units involving correlations among the neighbor frames based on the maximum likelihood parameter generation (MLPG) [25]. The MLPG is an algorithm that estimates the optimum sequence of features from static and dynamic features based on a maximum likelihood estimation. Because the CPCA features are complex-valued, we use the complex-valued gradient for parameter generation. Formulation of the complex extension of MLPG is presented as follows.

After training the CRBM, we estimated the optimum sequence of CPCA features $\hat{\boldsymbol{z}}_{1:T} \triangleq [\hat{\boldsymbol{z}}_1^H \ \hat{\boldsymbol{z}}_2^H \ \cdots \ \hat{\boldsymbol{z}}_T^H]^H$, where T is the number of frames of the test speech, from the encoded features (the expectations of hidden units) $\hat{\boldsymbol{h}}_t \triangleq \mathbb{E}[p(\boldsymbol{h}_t|\boldsymbol{z}_t)]$ that were calculated from the original concatenated features $\boldsymbol{Z}_{1:T} \triangleq [\boldsymbol{Z}_1^H \ \boldsymbol{Z}_2^H \ \cdots \ \boldsymbol{Z}_T^H]^H$ of the test speech. $\hat{\boldsymbol{z}}_{1:T}$ is the sequence that maximizes the conditional probability $p(\boldsymbol{Z}_{1:T}|\hat{\boldsymbol{h}}_{1:T})$, which is defined as:

$$\hat{\boldsymbol{z}}_{1:T} = \underset{\boldsymbol{z}_{1:T}}{\operatorname{argmax}}\, p(\boldsymbol{Z}_{1:T}|\hat{\boldsymbol{h}}_{1:T}). \tag{60}$$

Now introducing the weight matrix $\mathbf{S} \in \mathbb{R}^{VT \times PT}$ that is:

$$\mathbf{S} \triangleq [\mathbf{S}_1 \ \mathbf{S}_2 \ \cdots \mathbf{S}_T]^\top \otimes \boldsymbol{I}_{P \times P} \tag{61}$$

$$\mathbf{S}_t \triangleq [\boldsymbol{s}_t^{(1)} \ \boldsymbol{s}_t^{(2)}], \tag{62}$$

where $\boldsymbol{s}_t^{(1)} \in \mathbb{R}^T$ and $\boldsymbol{s}_t^{(2)} \in \mathbb{R}^T$ are the sparse vectors where only the t-th element is 1 otherwise 0, and where the $(t-1)$-th element has the value of -0.5 and the $(t+1)$-th element of 0.5 otherwise 0, respectively, the sequence can be rewritten as $\boldsymbol{Z}_{1:T} = \mathbf{S}\boldsymbol{z}_{1:T}$. Since the conditional probability in Eq.(22) has a single mode, the objective $\mathcal{Q} \triangleq \log p(\boldsymbol{Z}_{1:T}|\hat{\boldsymbol{h}}_{1:T}, \boldsymbol{\theta})$ can be calculated as:

$$\begin{aligned} Q = &- \boldsymbol{z}_{1:T}^\top \boldsymbol{S}^\top \operatorname{diag}(\tilde{\boldsymbol{q}}) \boldsymbol{S} \boldsymbol{z}_{1:T} - \boldsymbol{z}_{1:T}^\top \boldsymbol{S}^\top \operatorname{diag}(\tilde{\boldsymbol{p}}) \boldsymbol{S} \bar{\boldsymbol{z}}_{1:T} \\ &+ \boldsymbol{z}_{1:T}^\top \boldsymbol{S}^\top \boldsymbol{\mu}_{1:T} + K \end{aligned}, \tag{63}$$

where K is a constant that can be ignored in the estimation, $\tilde{\boldsymbol{x}}$ indicates a vector that put $\boldsymbol{x}$ for T times in a column, and

$$\boldsymbol{\mu}_{1:T} \triangleq [\boldsymbol{\mu}_1^H \ \boldsymbol{\mu}_2^H \ \cdots \ \boldsymbol{\mu}_T^H]^H \tag{64}$$

$$\boldsymbol{\mu}_t \triangleq \operatorname{diag}(\boldsymbol{p})(\boldsymbol{b} + \mathbf{W}\hat{\boldsymbol{h}}_t) + \operatorname{diag}(\boldsymbol{q})(\bar{\boldsymbol{b}} + \bar{\mathbf{W}}\hat{\boldsymbol{h}}_t). \tag{65}$$

We estimate the optimum sequence $\hat{\boldsymbol{z}}_{1:T}$ using a complex-valued gradient method in a way similar to that discussed in the previous section. Specifically, using the initial sequence as the frame-wise optima of the static features from:

$$\underset{\boldsymbol{Z}_t}{\operatorname{argmax}}\, p(\boldsymbol{Z}_t|\hat{\boldsymbol{h}}_t, \boldsymbol{\theta}) = \boldsymbol{b} + \mathbf{W}\hat{\boldsymbol{h}}_t, \forall t, \tag{66}$$

we iteratively update the sequence as:

$$\boldsymbol{z}_{1:T}^{(l+1)} \leftarrow \boldsymbol{z}_{1:T}^{(l)} + \boldsymbol{g}^{(l)} \left(\frac{\partial Q}{\partial \boldsymbol{z}_{1:T}}\right), \tag{67}$$

where $\frac{\partial Q}{\partial \boldsymbol{z}_{1:T}}$ indicates the Wirtinger derivative and can be calculated as:

$$\frac{\partial Q}{\partial \boldsymbol{z}_{1:T}} = -2\boldsymbol{S}^\top \operatorname{diag}(\tilde{\boldsymbol{q}}) \boldsymbol{S} \boldsymbol{z}_{1:T} - \boldsymbol{S}^\top \operatorname{diag}(\tilde{\boldsymbol{p}}) \boldsymbol{S} \bar{\boldsymbol{z}}_{1:T} + \boldsymbol{S}^\top \tilde{\boldsymbol{\mu}}. \tag{68}$$

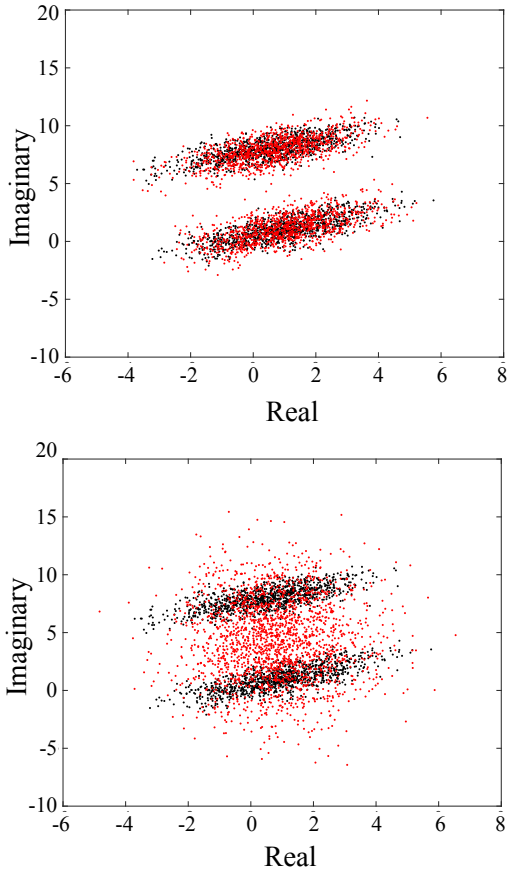


Fig. 4: Artificially created 1D complex-valued data (black dots) and random samples (red dots) generated from the trained models: the proposed CRBM (above) and the conventional RBM (below).

In our experiments, we employ the CSA for the gradient function $\boldsymbol{g}^{(l)}$ in the sequence generation.

VI. Experimental evaluation

A. Evaluation using artificial data

In order to confirm the effectiveness of our CRBM, we first conducted a simple experiment using one-dimensional complex-valued artificial data (the number of training data $N = 2000$). The artificially created data is illustrated in Fig. 4 as black dots, which have correlations between the real and imaginary parts. In this experiment, we compared the CRBM to a conventional RBM that has two real-valued visible units; one is for the real part, another is for the imaginary part. We trained both models with two hidden units using the steepest gradient ascent with a learning rate of 0.01, a momentum of 0.1, a batch size of 20, and a number of epochs as 200. After the training, we randomly generated samples from the models; the samples from the CRBM are shown as red dots on the top of Fig. 4, and the RBM are shown as red dots on the bottom of Fig. 4. As shown in Fig. 4, we can see that the proposed CRBM could represent the distribution of the complex-valued artificial data more accurately than the RBM. This is because the CRBM captures the relationships between the real and imaginary parts while the conventional RBM does not.

TABLE I: PESQ of the reconstructed speech from CPCA features.

P	**20**	**40**	**60**	**80**	**100**
PESQ	3.71	4.46	4.49	4.50	4.50

TABLE II: PESQ comparison of CPCA features and raw complex spectra as visible units in CRBM.

H	**CPCA** ($P = V = 40$)	**Complex spectra** ($V = 129$)
1k	**2.34**	2.30
2k	**2.60**	2.54
4k	**2.70**	2.66

B. Evaluation using speech data

Secondly, we conducted speech representation experiments using speech signals of 50 sentences (approx. 4.2 min) for training and another 53 for tests pronounced by a female announcer ("FTK") from the set "A" of the ATR speech corpora. The speech signals were downsampled from the original 20kHz to 16kHz due to speed-up of computation, and processed into 129-dimensional complex spectra using the short-time Fourier transform (STFT) with a window length of 256 and a hop size of 64 without pitch synchronous, followed by the CPCA to obtain complex-valued features. The total number of the training data was $64,438$. In order to decide how much to reduce the number of dimensions by the CPCA, we examined the perceptual evaluation of speech quality (PESQ) of the recovered signals using the inverse short-time Fourier transform (ISTFT) and the overlap-add method from the CPCA features by changing the number of dimensions P to 20, 40, 60, 80, and 100, as shown in Table I. Table I shows that the PESQ with $P = 40$ is similar to those with higher P. Furthermore, we also show the effectiveness of the CPCA features comparing with the case where raw complex-valued spectra were used as visible units of CRBM. Table II shows this comparison changing the number of hidden units H. The CPCA features outperformed the raw complex spectra regardless of H in the CRBM speech representation. Given these results, we used CPCA features with $P = 40$ as complex-valued visible units in the rest of our experiments in terms of sufficient quality and dimensional reduction.

1) Methods compared: We compared our model "CRBM" and its trajectory version "CRBM+T" with the RBM that feeds concatenated real-valued vectors of real and imaginary parts of the CPCA features ("RBM") and static and dynamic features ("RBM+T"). Another RBM-based method ("RBM+GL") we compared was trained using 40-dimensional real-valued features obtained by amplitude spectra followed by PCA as visible units and recovered speech signals using the Griffin-Lim algorithm [11]. These models were evaluated by changing the number of hidden units H to $1,000$, $2,000$, and $4,000$. The CRBMs were trained using the stochastic gradient method of 100-size mini-batches and 200 epochs with a learning rate $\alpha = 0.01$ for the CSA and $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$

TABLE III: Configuration of each method. P indicates the number of dimensions of CPCA (or PCA) features; that is, the number of static visible units. H is the number of hidden units.

H	**CRBM+T**	**CRBM**	**RBM+T**	**RBM**	**RBM+GL**
Base model	CRBM	CRBM	RBM	RBM	RBM
# of Params	$4PH + 10P + H$	$2PH + 5P + H$	$4PH + 8P + H$	$2PH + 4P + H$	$PH + 2P + H$
Feature	CPCA + Δ	CPCA	CPCA + Δ	CPCA	PCA
Optimizer	CAdam	CAdam	Adam	Adam	Adam
Phase Recovering	-	-	-	-	Griffin-Lim

TABLE IV: PESQ comparison of CAdam and CSA algorithm in CRBM.

H	**CAdam**	**CSA**
1k	**2.34**	2.29
2k	**2.60**	2.38
4k	**2.70**	2.34

TABLE V: PESQ of the CRBM and RBM methods when changing the number of hidden units (the leftmost column). The methods with the notation "+T" use trajectory estimation; otherwise, they use frame-wise estimation. "+GL" denotes the use of the Griffin-Lim algorithm.

H	**CRBM+T**	**CRBM**	**RBM+T**	**RBM**	**RBM+GL**
1k	**2.41**	2.34	2.39	2.30	2.33
2k	**2.72**	2.60	2.62	2.54	2.46
4k	**2.81**	2.70	2.66	2.54	2.39

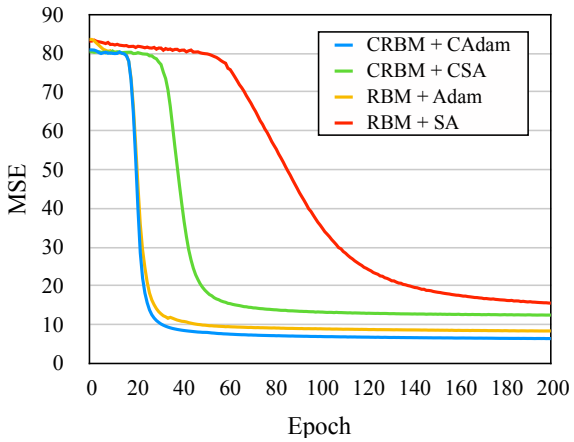


Fig. 5: MSE curve during the training of the CRBMs and the RBMs.

for the complex Adam (CAdam). We set the same parameters for the RBMs except for using the real-valued steepest ascent (SA) and Adam. For the gradient method to estimate the sequence in Eq. (67), we used the CSA (SA for the RBM method) of 100 epochs with a learning rate of 0.01. We summarized the configuration of each method in Table III.

We also compared our method to the traditional speech representation of cepstral ("CEP") and mel-cepstral ("MCEP") analysis. Both features were obtained by the FFT resolution of 256, which is the same as our method. The cepstral coefficients were 40 and recovered speech using the log magnitude approximation (LMA) filter [31]. From 20-dimensional mel-cepstral coefficients, we restored speech using the mel-log spectrum approximate (MLSA) filter [32]. Finally, we compared it to the WORLD [13] as a high-quality speech analysis-by-synthesis system. For fair comparison, the WORLD spectra were extracted by 4 msec. of frame shift, which is the same condition as our approach. We set the pitch range of the WORLD as $[71, 640]$ Hz.

2) Objective evaluation: Figure 5 shows the mean-squared error (MSE) between the training samples and the reconstructed samples (the expected values of visible units given hidden units that were calculated from the training samples), comparing the CRBM with CAdam ("CRBM+CAdam") to the counterparts. The MSE for the RBMs was calculated as the sum of the squared errors with respect to real and imaginary parts, which is comparable to the MSE of complex values for the CRBMs. As shown in Fig. 5, the CRBMs converged more quickly than the counterparts of RBMs, and the CAdam algorithm was considerably effective for the CSA. As shown in Table IV, where the CAdam and the CSA algorithms for training the CRBM are compared using PESQ from the test set, the CAdam was superior to the CSA on test set as well.

TABLE VI: Speech reconstruction performance (PESQ) of each method. The notations are the same as in Table V. The values after $\pm$ indicate the standard deviation of the PESQ.

Methods	**PESQ**
CRBM+T	$\mathbf{2.81 \pm 0.124}$
CRBM	2.70 ± 0.109
RBM+T	2.66 ± 0.122
RBM	$2.54 \pm 0.1.8$
RBM+GL	2.46 ± 0.123
MCEP	2.68 ± 0.299
CEP	2.54 ± 0.286
WORLD	$\mathbf{2.86 \pm 0.149}$

Table V illustrates the performance of the CRBM and RBM methods in the test set, showing that the proposed method outperformed the rest regardless of the number of hidden units. While the performance of the CRBMs with $H = 1,000$ was comparable to that of the RBMs with $H = 1,000$, the CRBMs with more hidden units highly improved the performance compared to the RBMs. The performance of the RBMs was not improved even when $H = 4,000$; that is, the performance was saturated. On the other hand, the CRBM with $H = 4,000$ was better than the CRBM with $H = 2,000$. We can say that the CRBMs are better able to represent complex-valued data than RBMs when a large amount of hidden units is given. We also conducted experiments to

TABLE VII: p-values of t-test between each comparison pair. The bold indicates no significant difference.

	CRBM+T	CRBM	RBM+T	RBM	RBM+GL	MCEP	CEP	WORLD
CRBM+T	–	**0.0731**	0.0009	0.0000	0.0000	0.0000	0.0000	**0.6852**
CRBM	–	–	**0.9208**	0.0001	0.0000	**0.1114**	0.0000	0.0001
RBM+T	–	–	–	0.0166	0.0000	**0.8077**	0.0000	0.0000
RBM	–	–	–	–	0.0005	**0.5654**	0.0204	0.0000
RBM+GL	–	–	–	–	–	0.0000	**0.9794**	0.0000
MCEP	–	–	–	–	–	–	0.0000	0.0000
CEP	–	–	–	–	–	–	–	0.0000
WORLD	–	–	–	–	–	–	–	–

TABLE VIII: Performance when changing the number of sentences used for training CRBM and RBM with trajectory modeling. The numbers indicate the PESQ.

# of sentences	CRBM+T	RBM+T
50	**2.81**	2.66
100	**2.85**	2.71
200	**2.94**	2.80

TABLE IX: PSNR [dB] of reconstructed spectra in the voiced range from the CRBM and RBM with respect to magnitude spectrum (MS) and phase difference (PD).

	MS	PD
CRBM	**39.8**	**7.04**
RBM	38.8	6.72

confirm the effectiveness of the CRBM against the RBM under the same amount of parameters. As Table III shows, the "CRBM" has roughly twice as many parameters as the RBM that models amplitude spectra ("RBM+GL"), while the number of parameters of the CRBM is almost the same as that of the RBM that models complex spectra ("RBM"). Therefore, we trained and evaluated the "RBM+GL" that has hidden units of 1978, 3954, and 7905, each of which corresponds to the "CRBM" with the number of hidden units as $H = 1,000$, $H = 2,000$, and $H = 4,000$ in terms of the same amount of parameters. The results were 2.39, 2.42, and 2.38 in the PESQ, respectively. These results also support the fact that the CRBM outperformed the RBM as the number of hidden units increased.

Table VI summarizes the performance of each method under their best conditions. All methods based on CRBM and RBM were trained using the CAdam and Adam algorithms. t-tests were applied to each pair of these methods, as Table VII shows the p-values. Interestingly, the CRBM with frame-wise modeling ("CRBM") significantly outperformed the RBM with trajectory modeling ("RBM+T") because the CRBM implicitly represents the phase information of complex-valued data, and the frame-wise features from the CRBM recovered speech sufficiently. Furthermore, the proposed trajectory modeling ("CRBM+T") improved accuracy by extracting the correlations of complex-valued features between adjacent frames and performed the best out of all the training-based methods, though there was no significant difference between "CRBM+T" and "CRBM." The performance of the proposed method is even comparable to that of the WORLD without significant difference, which is one of the highest-quality synthesis methods. Unlike traditional speech representation methods (mel-cepstrum, cepstum, and WORLD), the CRBMs directly extract latent features from arbitrary complex-valued features, which indicates that the CRBMs have high versatility with complex-valued data and can be applied to speech and to other signals. Furthermore, the training-based methods can improve their accuracies when fed more training data, as shown in Table VIII.

Table IX compares the peak signal-to-noise ratio (PSNR) of the CRBM and RBM in order to analyze which magnitude or phase in the reconstruction of the CRBM is actually effective. Because the instantaneous phase is sometimes unstable and meaningless, we use the phase difference (PD) instead. The PD is calculated from the unwrapped phase, followed by the differential with respect to frequency. The PSNR is similar to the signal-to-noise (SNR) except that it measures the distortion against the peak value. The PSNR is calculated as:

$$PSNR = 10 \log_{10} \frac{PEAK^2}{MSE}, \tag{69}$$

where $PEAK$ is the peak value and MSE is the mean-squared error between the estimated and the original spectra in terms of magnitude and phase difference. $PEAK$ for magnitude and phase evaluation were set as the maximum value of the magnitude spectra and 2π, respectively.

According to Table IX, the CRBM got 2.58% relative improvement points to the RBM in terms of magnitude and 4.76% relative improvement points to the RBM in terms of phase. We conducted a t-test and confirmed that there were both significant differences between the CRBM and RBM with the significance level of 5%. This demonstrates that the CRBM can effectively represent complex-valued data in particular with respect to phase.

3) Subjective evaluation: Finally, we conducted subjective experiments based on the mean opinion score (MOS) of 95 participants gathered through crowdsourcing. Each participant was asked to rank the synthesized and natural speech (NAT) on a 5-point scale (1: poor, 2: fair, 3: good, 4: very good, and 5: excellent) in terms of speech quality (naturalness). Because the MCEP, the CEP, and the WORLD are based on frame-wise synthesis, we used frame-wise estimation for the CRBM and the RBM in these experiments rather than trajectory estimation. Figure 6 shows the results of the subjective evaluation. As Fig. 6 shows, the CRBM performed the best out of all the methods except the WORLD. We also conducted pairwise t-tests for each combination and observed significant differences

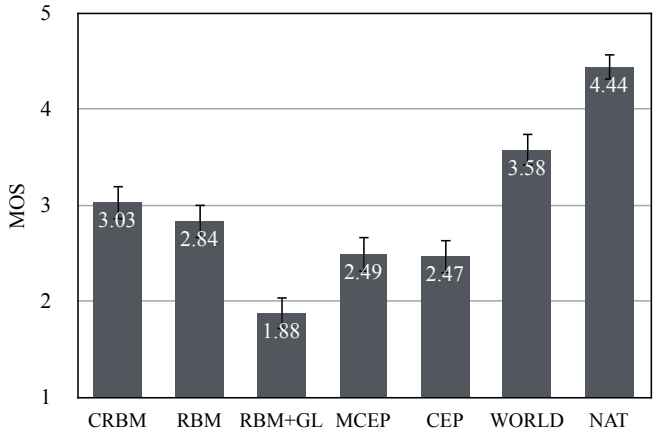


Fig. 6: Mean opinion score (MOS) for each method. The numbers in bars indicate the average of the MOS and the error bars indicate the confidential intervals.

with a 95% confidence for all pairs except for the difference between the CEP and MCEP.

VII. Conclusion

We proposed a "complex-valued RBM" (CRBM), a novel probabilistic model that extends an RBM in order to feed complex-valued data. This paper also includes its improved learning methods in modeling speech: the dimensionality-reduction of complex-valued data using CPCA, the CAdam learning algorithm to estimate complex-valued parameters more effectively, and the trajectory modeling and the generation method of complex-valued data based on MLPG. Experimental results showed the effectiveness of the proposed method with objective and subjective criteria compared to the other speech representation methods except the WORLD. Although the CRBM fell just one step short of the WORLD in terms of quality as a specialized speech representation method, the CRBM can be also used for representation of other signals, such as music, images, array signals, etc. We demonstrated the performance of the CRBM through basic experiments of analysis-synthesis reconstruction. Since our model can be used in more practical tasks, such as voice conversion, speech synthesis, and speech recognition, we will further investigate the high ability of the CRBM in such applications in the future. Furthermore, we showed the effectiveness of our method in terms of the speaker-dependent speech representation. However, the CRBM can be trained using the speech data of several speakers. We believe that the model is also effective to speaker independent cases, which will be investigated in the future.

We presented the CRBM in this paper as a very basic model and believe that the model can be easily extended. For example, we could define extensions by stacking two or more hidden layers like the deep Boltzmann machine [33] by adding connections from the previous to the current hidden/visible units like the recurrent temporal Boltzmann machine [34], or by changing the energy function to form the conditional probability of hidden units such as Gaussian distribution, complex normal distribution, etc. The deep extension can be also used as a pre-training method for complex neural networks [18]. Future work includes such extensions.

Acknowledgment

This work was partially supported by JST ACT-I, by MEXT KAKENHI Grant Numbers (15H01686, 16K16096, 16H06302,17H04687,18K18069), and by the Telecommunications Advancement Foundation Grant.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[2] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
[3] R. Salakhutdinov and G. E. Hinton, "Deep Boltzmann machines," *AISTATS*, pp. 448–455, 2009.
[4] A. Krizhevsky and G. E. Hinton, "Factored 3-way restricted boltzmann machines for modeling natural images," *Journal of Machine Learning Research*, 2010.
[5] K. Sohn, G. Zhou, C. Lee, and H. Lee, "Learning and Selecting Features Jointly with Point-wise Gated Boltzmann Machines," *ICML (2)*, 2013.
[6] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-Parallel Training in Voice Conversion Using an Adaptive Restricted Boltzmann Machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2032–2045, 2016.
[7] Y. Freund and D. Haussler, "Unsupervised learning of distributions of binary vectors using two layer networks," *Computer Research Laboratory*, pp. 912–919, 1994.
[8] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Advances in neural information processing systems*, 2008, pp. 873–880.
[9] S. Takaki, H. Kameoka, and J. Yamagishi, "Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis," in *Proc. Interspeech*, 2017, pp. 1128–1132.
[10] D. Palaz, R. Collobert *et al.*, "Analysis of CNN-based speech recognition system using raw speech as input," in *Proc. Interspeech*, 2017, pp. 11–15.
[11] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.
[12] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds1," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.
[13] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
[14] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *ArXiV (not peer reviewed)*, 2016.
[15] A. v. d. Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," *ArXiV (not peer reviewed)*, 2017.
[16] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Proc. Interspeech*, 2014, pp. 2514–2518.
[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
[18] I. Nemoto and T. Kono, "Complex neural networks," *Systems and computers in Japan*, vol. 23, no. 8, pp. 75–84, 1992.
[19] R. S. Zemel, C. Williams, and M. C. Mozer, "Lending direction to neural networks," *Neural Networks*, vol. 8, no. 4, pp. 503–512, 1995.
[20] H. Kameoka, N. Ono, and K. Kashino, "Complex NMF: A new sparse representation for acoustic signals," *Proc. ICASSP 2009*, pp. 3437–3440, 2009.

[21] R. Maia, M. Akamine, and M. J. Gales, "Complex cepstrum as phase information in statistical parametric speech synthesis," in *Proc. ICASSP 2012*, 2012, pp. 4581–4584.
[22] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 38, 2014.
[23] T. Nakashika, S. Takaki, and J. Yamagishi, "Complex-valued restricted Boltzmann machine for direct learning of frequency spectra," in *Proc. Interspeech*, 2017, pp. 4021–4025.
[24] J. Horel, "Complex principal component analysis: Theory and examples," *Journal of climate and Applied Meteorology*, vol. 23, no. 12, pp. 1660–1673, 1984.
[25] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, 2000, pp. 1315–1318.
[26] K. Cho, A. Ilin, and T. Raiko, "Improved learning of Gaussian-Bernoulli restricted Boltzmann machines," in *Proc. ICANN*. Springer, 2011, pp. 10–17.
[27] B. Picinbono, "Second-order complex random vectors and normal distributions," *IEEE Transactions on Signal Processing*, vol. 44, no. 10, pp. 2637–2640, 1996.
[28] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proceedings F-Communications, Radar and Signal Processing*, vol. 130, no. 1, pp. 11–16, 1983.
[29] H. Zhang and D. P. Mandic, "Is a complex-valued stepsize advantageous in complex-valued gradient learning algorithms?" *IEEE transactions on neural networks ...*, vol. 27, no. 12, pp. 2730–2735, 2016.
[30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.
[31] S. Imai, "Speech analysis synthesis system using the log magnitude approximation filter," *Trans. Inst. Electron. Commun. Eng. Jpn.*, vol. 61, pp. 527–534, 1978.
[32] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.
[33] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," in *Artificial Intelligence and Statistics*, 2009, pp. 448–455.
[34] I. Sutskever, G. E. Hinton, and G. W. Taylor, "The recurrent temporal restricted Boltzmann machine," in *Advances in Neural Information Processing Systems*, 2009, pp. 1601–1608.

Toru Nakashika received his B.E. and M.E. degrees in computer science from Kobe University in 2009 and 2011, respectively. In the summer in 2010, he was a student researcher at IBM Research, Tokyo Research Laboratory. From September 2011 to August 2012, he was a visiting researcher in the image group at INSA Lyon in France. In that year, he continued his research as a doctoral student at Kobe University and received his Dr.Eng. degree in computer science in 2014. He was an Assistant Professor at Kobe University until April 2015. He is currently an Assistant Professor at the University of Electro-Communications in Chofu, Japan. He received the IEICE ISS Young Researcher's Award in Speech Field in 2013, the IPSJ Ongaku Symposium Excellent Paper Award in 2016, and the Awaya Award from the Acoustical Society of Japan in 2018. He is a member of IEEE, ISCA, IEICE, and ASJ.

Shinji Takaki received a B.E. degree in computer science and received an M.E. and a Ph.D. degree in scientific and engineering simulation from the Nagoya Institute of Technology, Nagoya, Japan in 2009, 2011, and 2014, respectively. From September 2013 to January 2014, he was a visiting researcher at the University of Edinburgh University. Since April 2014, he has been a project researcher at the National Institute of Informatics in Japan. His research interests include statistical machine learning and speech synthesis. He is a member of the Acoustical Society of Japan and the Information Processing Society of Japan.

Junichi Yamagishi (SM'13) is an associate professor of the National Institute of Informatics in Japan. He is also a senior research fellow at the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, UK. He was awarded a Ph.D. by the Tokyo Institute of Technology in 2006 for a thesis that pioneered speaker-adaptive speech synthesis and was awarded the Tejima Prize for the best Ph.D. thesis of the Tokyo Institute of Technology in 2007. Since 2006, he has authored and co-authored over 100 refereed papers in international journals and conferences. He was awarded the Itakura Prize from the Acoustic Society of Japan, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan, the Young Scientists ' Prize from the Minister of Education, Science, and Technology, and the JSPS prize in 2010, 2013, 2014, and 2016, respectively.

He was one of organizers for special sessions on "Spoofing and Countermeasures for Automatic Speaker Verification" at Interspeech 2013, "ASVspoof evaluation" at Interspeech 2015, "Voice conversion challenge 2016" at Interspeech 2016, and "2nd ASVspoof evaluation" at Interspeech 2017. He has been a member of the Speech & Language Technical Committee (SLTC). He served as an Associate Editor of the IEEE/ACM Transactions on Audio, Speech, and Language Processing and as a Lead Guest Editor for the IEEE Journal of Selected Topics in Signal Processing (JSTSP) special issue on Spoofing and Countermeasures for Automatic Speaker Verification.